Retrospective analysis of main and joint effects in genetic association studies of human complex traits

Qihua Tan[1,2], Lene Christiansen[1,2], Charlotte Brasch-Andersen[2,3], Jing Hua Zhao[4], Shuxia Li[1], Torben A. Kruse[2], Kaare Christensen[1]

1. Epidemiology, Institute of Public Health, University of Southern Denmark, Denmark
2. Department of Biochemistry, Pharmacology and Genetics, Odense University Hospital, Denmark
3. Pharmacology, Clinical Institute, University of Southern Denmark, Denmark
4. MRC Epidemiology Unit, The Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK

BMC Genet 2007, **8**:70

Outline

- A run-through of the paper
  - Summary
  - Background
  - Models
  - Results
  - Discussion
- General remarks
  - Implications
  - Broad framework
  - Other aspects
- Comments? suggestions?

Summary

Multifactorial diseases involve complex interactions between environmental factors and genes, which requires efficient statistical tools to identify the genetic and environmental variants that affect the risk of disease. This paper introduces a retrospective polytomous logistic regression (RPLR) model to measure both the main and joint effects in genetic association studies of human complex traits both discrete and continuous, whereby combinations of genotypes at the two interacting loci or of environmental exposure and genotypes at one locus are treated as nominal outcomes and whose proportions are modeled as a function of the disease trait assigning both main and interaction effects and with no assumption on normality of the trait distribution. Performance of the method in detecting joint or interaction effect is compared with the case-only model.

The model is applied to data on catelase -262C/T promoter polymorphism and aging phenotypes and detected significant effects for age-group and allele T on individual's cognitive functioning and produces consistent results in estimating the joint effect as compared to the popular case-only model.

The RPLR model is a convenient tool for assessing both main and joint effects in genetic association studies of categorical or continuous traits involving genetic and non-genetic factors.

Background

- Genetic association study is a powerful tool to identify genetic and non-genetic factors contributing to aetiology of common diseases
- Retrospective methods have been proposed for both single locus and multi-locus, e.g. Tan et al. Ann Hum Genet 2003, **67**:598-607; Tan et al. Genet Res 2005, **86**:223-31
- Transmission disequilibrium test (TDT), e.g. Waldman et al. Ann Hum Genet 1999, **63**:329-40
- Case-control design, e.g. Epstein MP, Satten GA. Am J Hum Genet 2003, **73**:316-29; Weinberg CR, Umbach DM. Am J Epidemiol 2000, **152**:197-203; Tan et al. Genet Res 2005, **86**:223-31
- Case-only design, e.g. Khoury & Flanners. Am J Epidemiol 1996, **144**:207-13

Models

Let G = genetic variant, 0=non-carriers, 1=carriers

E = environment exposure, 0=non-exposed, 1=exposed

x = trait value

Then, we have

Logit $P[G=1,E=1|x]=a_G I+b_E J+(b_G I+b_E J+b_{GxE} IJ)x$

Where *a*=intercept (nuisance parameter), *b*=slope

We can write out $P[G=1,E=1|x]$, $P[G=0,E=0|x]$, $P[G=1,E=0|x]$ and therefore $RR_G(x)=exp(a_G+b_G x)$, $RRR_G(x_2:x_1)=exp[b_G(x_2-x_1)]$

The log-likelihood function is $l=\sum_{k=1}^{N}\sum_{i,j=0}^{1} I(G_k=I,E_k=J)p(a_G,a_E,b_G,b_E,b_{G\times E})$

Results

- Simulation **(Table 1)** according to $\gamma_i = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 G * E + e_i,\ e_i \sim N(0,1)$
  - Null hypothesis
  - Alternative hypothesis
  - P(x)=0.2, P(E)=0.3,
- Real data (*N=789*) from $\beta_0 = 0.1$ Christiansen et al. J Gerontol A Biol Sci 2004, **59**:8886-9
  - G = catalase -262C/T promotor polymorphism and aging
  - E = age group
- Results shown in **Table 2**. The likelihood function based on continuous trait is preferred due to larger value

Table 1. Power and empirical type I error rate for α=0.05

Sample size	Power			Type-I error
	β=-0.65	β=-0.95	β=-1.20	β=0
150	0.348	0.642	0.780	0.052
200	0.540	0.668	0.894	0.048
250	0.604	0.852	0.898	0.054
300	0.664	0.884	0.960	0.046
400	0.760	0.948	1.000	0.050
600	0.876	1.000	1.000	0.052

Table 2. Parameter estimates for main and interaction effects on cognitive score by the logistic regression model

				Risk			*l**
	Slop	SE	p-value	RRR	95%	CI	
Continuous							
Age-group	-0.463	0.036	0	0.63	0.587	0.676	
Allele T	-0.054	0.026	0.037	0.948	0.901	0.997	
Joint effect	0.079	0.037	0.033	1.083	1.006	1.164	
							-862.66
Dichotomous							
Age-group	-2.822	0.253	0	0.06	0.036	0.098	
Allele T	-0.411	0.202	0.041	0.663	0.446	0.984	
Joint effect	0.87	0.333	0.009	2.386	1.241	4.588	
							-928.45

**l*=maximum log-likelihood

Discussion

- A distinct feature of the model is trait is treated as an independent variable to allow for both categorical and continuous types and no assumption about normality is required (Cordell H. Hum Mol Genet 2002, **11**:2463-8; Hahn et al. Bioinformatics 2003, **19**:376-82).
- Supplementary materials showed that for the case of binary outcome, when the disease is rare, the relative risk estimate is comparable to that from a prospective model.
- It requires interaction variants be independent, otherwise a haplotype analysis model is required (Epstein MP, Satten GA. Am J Hum Genet 2003, **73**:1316-29; Tan et al. Genet Res 2005, **86**:223-31).
- It is equally applicable to study of any main and joint effects models

General remarks

- Philosophical implication
  - A paradigm of Cause ← Outcome
  - Proposed and discussed in DiOGenes analysis
- Broad framework
  - Prentice RL, Pyke R. Biometrika 1979, **66**:403-11.
  - In this context, the full retrospective likelihood is proportional to the retrospective likelihood of gene conditional on environmental factors and disease multiplied by the prospective likelihood of disease given genetic factors (Kwee et al. Genet Epidemiol 2007, **31**:73-90). i.e., $L = P[G,E|D] = P[G|E,D]P[E|D]$
- Other aspects
  - Power and bias were examined by Satten GA, Epstein MP, Genet Epidemiol 2004, **27**:192-201
  - Modification of power calculation sample size as accorded to rare and smaller effect size, e.g., InterAct. Gauderman WJ, Morrison JM. QUANTO: A computer program for power and sample size calculations for genetic-epidemiology studies, http://hydra.usc.edu/gxe.
  - The inclusion of covariates, e.g., Zou GY. Ann Hum Genet 2006, **70**:262-72
  - To relax the independence assumption. Shin J-H, McNeney B, Graham J. Stat Appl Genet Mol Biol 2007, **6:**13

More details from Kwee et al.

- A prospective likelihood approach by Lake et al. Hum Hered 2003, **55**:56-65 may suffer with respect to retrospective approach (Epstein & Satten 2004). Profile likelihood approaches by Lin et al. Genet Epidemiol 2005, 29:299-312; Spinka et al. Genet Epidemiol 2005, **29**:108-127 require estimating absolute risk of disease from case-control data
- Consider (a) rare disease and (b) haplotype-environment independence
- This leads to likelihood-based inference without specifying the distribution of environment covariates in the sample. Following the full likelihood specification, it is shown that $P[E|D]$ contained no information on haplotype and haplotype-environment interaction parameters. Furthermore, case-only study relies on information from $P[G|E,D=1]$
- Simulation studies showed that (with a realistic sample size) under a recessive model both the full retrospective and the prospective approaches yielded flawed results and occasional convergence problem. However, the multiplicative and dominant models give reasonable estimate.
- (To be) implemented in CHAPLIN for case-control haplotype analysis

Comments? Suggestions?

- Is it viable in your opinion?
- Further work?